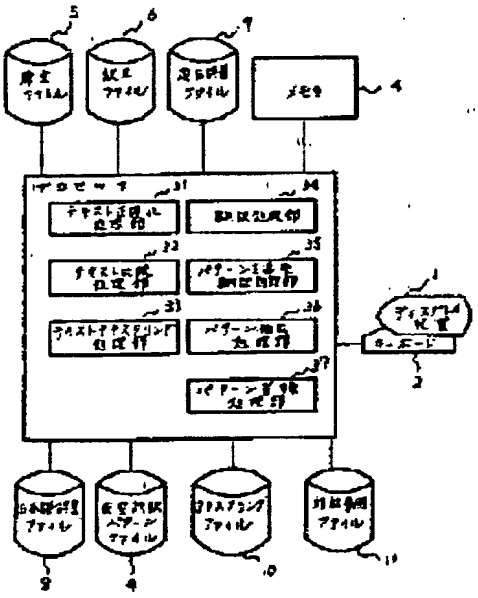
METHOD AND DEVICE FOR EXAMPLE REUSE TYPE TRANSLATION

Patent number:	JP4160473
Publication date:	1992-06-03
Inventor:	YAMANO FUMIYUKI
Applicant:	HITACHI LTD
Classification:	
- international:	(IPC1-7): G06F15/38
- european:	
Application number:	JP19900284240 19901024
Priority number(s):	JP19900284240 19901024

Abstract of **JP4160473**

PURPOSE:To improve the translation accuracy and operation efficiency by holding patterns of similar sentences together with translation, side by side, and applying a translating process system which principally uses patterns for similar sentences. CONSTITUTION:A text clustering process part 33 clusters a document to be translated or translation document by similar sentences and stores it in a clustering file 10. The file 10 contains clustering numbers, source sentence numbers, and translation numbers in cluster units so that they are made correspond to one another. Further, model translation pattern numbers for access to a model translation pattern file 9 are further stored, cluster by cluster. Then a model pattern in translation correspondence form is extracted from the clustered sentences in a translation example base, a similar translation model pattern is retrieved for an input sentence, and the retrieved translation model pattern is utilized to translate only nonsimilar elements in the sentence partially, thereby composing the translation of the whole sentence.



⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平4−160473

⑤1Int. Cl.⁵	識別記号	庁内整理番号	⑬公開　平成4年(1992)6月3日
G 06 F 15/38		Q 9194−5L	
		J 9194−5L	
		T 9194−5L	

審査請求　未請求　請求項の数 18　(全20頁)

⑤4発明の名称　事例再利用型翻訳方法および装置

⑳特　願　平2−284240
㉒出　願　平2(1990)10月24日

⑦2発明者　山野　文行　神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
⑦1出願人　株式会社日立製作所　東京都千代田区神田駿河台4丁目6番地
⑦4代理人　弁理士　小川　勝男　外1名

明　細　書

1．発明の名称

事例再利用型翻訳方法および装置

2．特許請求の範囲

1．第1言語から第2言語への言語間の翻訳を行なう翻訳処理装置において、

第1言語で記述されたテキストから構成される翻訳対象文書について類似の文毎にクラスタリングする手段と、翻訳済みのテキストに対して類似の文毎にクラスタリングを行い、対訳形式で対訳事例ベースとして管理する手段と、前記の対訳事例ベースのクラスタリングされた文から対訳形式の典型パターン（以下、対訳典型パターンと云う）を抽出する手段と、入力文に対して類似の対訳典型パターンを検索し、検索した対訳典型パターンを利用して、文中の類似しない要素のみ部分的に翻訳処理を行ない、文全体の訳文を合成する手段を有することを特徴とする翻訳処理装置。

2．上記の対象事例ベースの作成方法として、翻訳対象文書と類似な対訳文書を予め類似度によりクラスタリングしておき、併せて典型対訳パターンを抽出する手段と、翻訳対象文書の翻訳校正時に類似の対訳文および典型対訳パターンを抽出・表示する手段を有することを特徴とする特許請求範囲第1項記載の翻訳処理装置。

3．特許請求の範囲第1項記載の翻訳処理装置において、

翻訳対象文書を文書中のテキストの出現順序に従って翻訳する際に、当該テキストがクラスタリングされた類似の文を含むかどうかチェックする手段と、当該テキストが類似のクラスタリング中に未翻訳であるテキストが持つ場合、当該テキストの翻訳と並行して類似のクラスタリング中のすべてのテキスト翻訳校正処理を行なう手段を有することを特徴とする翻訳処理装置。

4．特許請求の範囲第1項記載の翻訳処理装置において、

翻訳対象文書をあらかじめクラスタリングす

る手段と、各クラスタリング中の代表となる典型文を抽出し、典型文のみ翻訳・編集した後典型対訳パターンを抽出する手段と、典型文との類似の文の翻訳はパターン主導型の翻訳方法により翻訳を行ない、類似文を一切持たない文は別の翻訳処理方法を適用することを特徴とする翻訳処理装置。

5．上記の翻訳文の校正処理において、典型対象パターンを利用して翻訳・校正する際に、典型対訳パターン中のデータとの共通部と異なり部とを識別表示することを特徴とする特許請求の範囲第1項記載の翻訳処理装置。

6．上記の翻訳・校正の際に、原文と訳文の対訳表示画面上で、一文毎のテキストに対応して類似文や典型対訳パターンの存在の有無を識別表示することを特徴とする特許請求の範囲第1項記載の翻訳処理装置。

7．上記の翻訳・校正の際の原文と訳文のそれぞれの表示において、典型対訳パターンに対応した共通部を書き換え不可能とし、異なり部を書き換え可能として表示する手段を有することを特徴とする特許請求範囲第1項記載の翻訳処理装置。

8．上記原文と訳文の表示において、典型対訳パターンに対応した共通部を書き換え不可能とし、異なり部を書き換え可能とした場合、共通部分の書替え変更操作を設けることにより、ワンクッションおいて共通部分の変更を行なう手段を有することを特徴とする特許請求範囲第7項記載の翻訳処理装置。

9．テキストの類似な関係によるクラスタリング処理において、クラスタリング対象のテキスト中の単語を、数字列表記の単語，固有名詞，人名，会社名に該当する単語、さらに単語の属性などの特定の条件に合致する単語列を別表記の単語列に置き換えた後、テキスト間の類似度を検出する手段を有することを特徴とするテキストのクラスタリング方法。

10．上記特許請求の範囲第2項もしくは第3項記載の類似の対訳文および典型対訳パターンを表示する際に、表示データが複数存在する場合は類似度の大きい順に表示する手段を有することを特徴とする特許請求範囲第1項記載の翻訳処理装置。

11．上記複数の類似文を表示する際に、類似文同士で完全一致する類似文は表示しない手段を有することを特徴とする特許請求範囲第2項もしくは第3項もしくは第10項記載の翻訳処理装置。

12．上記複数の類似文を表示する際に、原文と完全一致する類似文は表示しない手段を有することを特徴とする特許請求範囲第2項もしくは第3項もしくは第10項記載の翻訳処理装置。

13．原文と訳文からなる対訳事例データに対して、原文のクラスタリングと訳文のクラスタリングの両方を行ない、クラスタリング数の近似するものを典型類似事例として扱う手段を有することを特徴とする対訳文のクラスタリング方法。

14．2つのテキスト間の類似度を算出するステップと二項関係のうちの少なくとも1つの要素が共通する関係同士をグループ化するステップからなることを特徴とするテキストのクラスタリング方法。

15．テキストの類似関係によりクラスタリング手段と、クラスタリングされたテキストからの典型対訳パターン抽出手段を有することにより、典型対訳パターンを原文と訳文の対訳型式の事例ベースから自動的に抽出することを特徴とする対訳テキストパターンの自動学習方法。

16．上記の典型対訳パターンの抽出方法において、新たに入力したテキストの翻訳・編集結果を逐次蓄積し、併せて既存のクラスタリングデータとのクランスタリングを行ない、クラスタリング文数がある閾値以上になったら、典型対訳パターンを抽出することを特徴とする対訳テキストパターンの動的な自動学習方法。

17．典型対訳パターン間の複数のパターンに対して、類似なパターン毎にクラスタリングを行なう手段と、クランスタリングされたパターン毎にパターンを統合する手段を有することを特徴

とする典型対訳パターンの再整理・統合方法。

18．2か国語の対訳関係を、3か国語以上の関係に拡張し、nか国語の対応関係を保持する対訳事例ベースおよび典型対訳パターンを有することを特徴とする特許請求の範囲第1項記載の翻訳処理装置。

3．発明の詳細な説明

〔産業上の利用分野〕

本発明は、ある言語で記述されたテキストを編集する文書編集装置、および、第1言語で記述されたテキストを第2言語で記述されたテキストに変換する言語間の翻訳処理装置に関し、特に類似の文を多く含む文を効率的に編集したり翻訳するのに好適な処理方式に関するものである。

〔従来の技術〕

従来より、第1言語から第2言語への翻訳処理装置においては、処理の対象を一文単位に限定しており、複数の文から構成されるテキストを翻訳するような場合でも、テキストから一文ずつ取り出して翻訳処理を行ない、その翻訳結果を連接して翻訳テキストを得るという方式が一般に採られている。この種の翻訳処理に関連する発明としては、特開昭58－40684 号（「自然言語間の自動翻訳方式」）や特開昭59－121574号（「翻訳処理方式」）等が挙げられる。

〔発明が解決しようとする課題〕

翻訳処理装置で扱う文書は、一般に複数の文から構成されるものが多い。内容的にまとまった文書であれば、類似の文を多用することが一般的である。さらに、扱う文書の分野や文書の種類が限定されると、同一文や類似文が数多く含まれる傾向にある。この代表例として、マニュアルを挙げることができる。

従来の翻訳処理装置では、類似の文を多く含む文書を翻訳する場合、既に翻訳済みの文と同一または類似な文が再度入力されても、新たに翻訳処理を駆動するものである。その結果、同一の文を何度も翻訳するといった処理時間の無駄や、類似文の翻訳結果が均一にならないといった問題点があった。

本発明の目的は、このような従来の問題点を改善し、類似の文を効率よく管理し、類似文のパターンを対訳形式で保持し、類似文に対してはパターン主体の翻訳処理方式を適用することにより、重複した翻訳処理を回避し、類似文の翻訳精度を均一に維持する、翻訳作業効率の高い翻訳処理装置を提供することにある。さらには、翻訳処理のみならず、文書を管理・維持していく上で、類似の文を効率よく管理し、類似文のパターンを保持し、類似文に対しては均一の表現に統一する文書編集処理装置を提供することも本発明の目的である。

〔課題を解決するための手段〕

上記の問題点を解決するため、本発明の機械翻訳方式は、

予め、翻訳対象文書について類似の文毎にクラスタリングする手段と、既に翻訳した文書に対して類似の文毎にクラスタリングを行い、対訳形式で対訳事例ベースとして管理する手段と、前記の対訳事例ベースのクラスタリングされた文から対訳形式の典型パターンを抽出する手段と、入力文に対して類似の対訳典型パターンを検索し、検索した対訳典型パターンを利用して、文中の類似しない要素のみ部分的に翻訳処理を行ない、文全体の訳文を合成する手段を有することを特徴とする。

さらに、本発明の文書処理装置では、予め編集対象文書について類似の文毎にクラスタリングする手段と、類似文に関しては類似文毎に管理する手段と、文書中の文を編集する際に、類似文を有する文かどうかをチェックする手段と、類似文を含む文を編集した場合は、類似文を抽出・表示し、編集者に類似文の編集の必要性をチェックさせる手段を有することを特徴とする。

〔作用〕

上記手段により、文発明を適用いた翻訳処理装置では、既に翻訳した文書の対訳を利用することにより、類似文を翻訳する場合、類似典型パターンと異なる要素の部分のみ翻訳すればよく、翻訳処理を軽減することができる。さらに、入力文書に対しても予め類似文毎にクラスタリングしてお

くことにより入力文全てを機械翻訳処理の対象とする必要がなくなり、訳翻処理効率が向上する。また、従来方式では、類似文でも構成要素が微妙に異なっただけで入力文の構文解析結果が微妙に異なったり、解析処理に失敗したりといった問題が発生していたが、本発明により、類似の文の翻訳精度は均一に保てるというメリットも生じる。

一方、本発明を適用した文書処理装置においては、文書の編集操作時に、文書中の類似文を確認することが容易になるため、マニュアルなどの編集操作において、表現のバラツキを防止し、表現上統一された文書の校正処理効率が向上する。

さらに、典型対訳パターンの抽出手段により、翻訳事例からパターンの自動学習が可能となる。すなわち、翻訳結果を再利用することにより翻訳処理のためのパターンデータを逐次自動的に増補していくことが容易に実現できるため、翻訳装置の利用者は翻訳規則の調整を行わずとも翻訳精度を向上させることが可能となる。

〔実施例〕

実施例１

以下、本発明を実施例を参照して詳細に説明する。一実施例として、英語から日本語への翻訳を行なう翻訳処理装置について述べる。

第１図は、本発明の一実施例を示すブロック図を表わす。図中、１はディスプレイ等の出力装置、２はキーボード等の入力装置、３はプロセッサ、４はメモリ、５は原文ファイル、６は訳文ファイル、７は英日辞書ファイル、８は日本語辞書ファイル、９は囲型対訳パターンファイル、１０はクラスタリングファイル、１１は対訳事例ファイルをそれぞれ示す。プロセッサ３は、さらに、テキスト正規化処理部３１，テキスト比較処理部３２，テキストクラスタリング処理部３３，翻訳処理部３４，パターン主導型翻訳処理部３５，パターン抽出処理部３６，パターン登録処理部３７から構成される。

次に、本発明による翻訳処理の流れを第２図に示すフローに従って具体的に説明する。

（２０１）英文ファイル５に格納されている翻訳対象英文について、全ての英文テキスト同士の類似度を算出したのち、英文テキストのクラスタリングを行なう。

ここで、２つのテキストの類似度の算出方法について説明しておく。基本的な考え方は、テキストの構成要素が出現順序で一致するかどうかの相関を調べることにより類似度を求める。

すなわち、第５図（ａ）のグラフに示すように、２つのテキストの構成要素（英語の場合は単語，日本語の場合は文節とすればよい）を抽出し、それぞれＸ軸，Ｙ軸に出現順序にしたがって配置する。続いて、Ｘ軸とＹ軸で構成要素が一致する座標を＊でプロットする。プロットした点について、Ｙ＝Ｘ＋α（αは変動してよい）の関係を満たし、プロット点をたどった場合、右上がりの線になるような最長の線を求める。ここで、順序関係が隣同士の構成要素にあるプロット点がｎ個連続して出現する場合、連接一致距離を（ｎ－１）と定義する。図中、Ｗａ１からＷａ３までは、連続一致距離が２とばる。この連続一致距離と、＊でプロットされた点を辿った軌跡が右上がりの線となる構成要素の一致個数を用いて、２つのテキストの類似度を次式で算出する。

$$\text{類似度}=\frac{(\text{連接一致距離}\times 2+\text{一致構成要素の個数}-\text{不一致構成要素の個数})}{\text{構成要素の総数}}$$

但し、テキストの末尾の句読点は、類似度算出の対象外とする。

実際に類似度を算出した例を、第５図（ｂ）と（ｃ）に示す。第５図（ｂ）の３つの英文は、上記の算出式に従えば、お互いに類似度が０．８６になる。一方、第５図（ｃ）では、類似度は0.71になる。

以下、第４図に示すフローに従って、類似度の評価方法について説明する。

（４０１）翻訳対象の全ての英文テキストについて、テキストの構成要素に分割する。ここでは、英文テキストが対象であるので単語間の空白や句読点をチェックすることにより単語毎に分割していけばよい。日本文テキストの場合は、日本語辞書ファイル４中の自立語や付属語の情報を利用し

て文節単位に分割すればよい。日本語テキストからの文節抽出方法については、長尾編「言語の機械処理」（１９８４年刊三省堂発行）のｐｐ.６１－８１に記載された方法を利用することが可能であり、説明は省略する。

（４０２）テキストの構成要素を正規化する。正規化処理とは、次のような構成要素列の置き換え、削除を行なう。ここでは、英文テキストを対象にした正規化処理について説明する。

(1) 数字から構成される単語を、NNNNNNに置換する。

(2) 大文字で始まる単語を、固有名詞とみなして、PPPPPPに置換する。

(3) NNNNNNが連続して現れる場合、１つの構成要素に縮退させる。

(4) PPPPPPが連続して現れる場合、１つの構成要素に縮退させる。

さらに、正規化処理の一環として、英日辞書ファイル７を利用して形態素処理を行うことにより、活用変化語を原形に変換することも考えられる。形態素処理方法については、特開昭58－40684 号に開示された方法を利用することが可能であり、説明は省略する。

（４０３）正規化処理を施したテキストについて順次２つのテキストを抽出し、テキスト間の類似度を算出する。

（４０４）類似度がある閾値（例えば、０.５）より大きい場合には、類似文によるクラスタリング処理を行なうべく、（４０５）の処理へ移行する。類似度な閾値より小さければ、（４０６）の処理へ移行する。

（４０５）類似度のある２つの英文テキストの文番号をペアにしてメモリ４に一時退避しておく。

（４０６）類似度の算出を全てのテキストの組み合わせについて行なったかどうかチェックし、未完であれば（４０３）、完了であれば(４０７）の処理へ移行する。

（４０７）メモリ４中の類似の関係にある英文テキストの文番号の対（二項関係）を調べ、２つの二項関係から部分的に一致する文番号があれば、同一のクラスタにセットしていく。この操作を有限回繰り返すことにより、お互いに類似関係のないクラスタが抽出できる。その結果をクラスタリングファイル１０に格納する。

ここで、第３図を用いて、クラスタリングファイル１０と他のファイルの関係を説明する。

クラスタリングファイル１０には、クラスタリングされた英文テキストの各クラスタ単位に、クラスタリング番号と複数個の英文テキストの文番号（原文番号）を蓄積しておく。併せて、個々の英文テキストに対応する訳文の文番号(訳文番号)を原文番号と対応して蓄積していく。原文番号，訳文番号は、それぞれ、原文ファイル５と訳文ファイル６に蓄積されている原文テキストと訳文テキストをアクセスするための文番号に対応する。さらに、クラスタ毎に、典型対訳パターンファイル９へのアクセス用の典型対訳パターン番号を格納しておく。

原文ファイル５には、原文テキストに対応して、文番号と登録日時、さらに、テキストが更新された際に記入する更新日時と、クラスタリングファイル中の対応するクラスタリング番号を格納しておく。訳文ファイル６には、原文ファイル５中の原文に対応する文番号と同一の文番号をアクセスキーとして、訳文テキスト，更新日時、さらに、クラスタリングファイル中の対応するクラスタリング番号を格納しておく。

テキスト間の類似度を利用して英文テキストのクラスタリング処理により、クラスタリングファイル１０には、クラスタ毎に類似の英文テキストの文番号が蓄積されている。これらの類似の英文中、出現順序が最初の英文テキストをクラスタの典型テキストとする。

（２０２）原文ファイル５から翻訳対象の英文テキストを出現順に抽出する。

（２０３）英文テキストが、類似文を持つかどうかチェックし、持てば（２０４）、持たなければ（２０７）の処理へ移行する。ここで、チェック方法としては、原文ファイル５中のクラスタリング番号がセットされているかどうかをチェック

すればよい。クラスタリング番号がセットされていないということは、原文テキスト中に類似の文が存在しないことを示す。類似の文が存在しない場合は、従来通りの翻訳処理を適用すればよい。

（２０４）英文テキストが類似文を持つので、典型対訳ファイル９をアクセスし、対応する典型訳文パターンが存在するかどうかチェックする。典型訳文パターンがセットされていれば（２０５）へ、セットされていなければ（２０８）の処理へそれぞれ移行する。

（２０５）典型対訳ファイル９に登録してある典型原文パターンと典型訳文パターンの対応関係を利用して、パターン主導型の翻訳処理を行なう。

ここで、パターン主導型の翻訳処理について、第７図に示す処理フローに従って説明する。

（７０１）翻訳対象の英文テキストと典型原文パターンの構成要素（この場合は単語とみなしてよい）を比較し、構成要素の共通部分と異なり部分を区別する。異なり部分をその出現順序に従って、順次、変数化部１，変数化部２，変数化部３，…変数化部ｎ，…のように対応付けしておく。

一例として第８図に示すような英文テキスト、"This character set has a different mapping from ＡＳＣＩＩ code."を翻訳する場合、典型原文パターンとして、"This !N set has a differens mapping fors !N code."が登録されていれば、"character"が変数化部１で最初の!Nに対応付けされ、"ＡＳＣＩＩ"が変数化部２で２番目の!Nに対応付けされる。

（７０２）変数化部として抽出された単語列を、順次、翻訳する。その際、原文パターン中の変数化部に対応する要素として、名詞句であれば!N、動詞句であれば!VP のように構文要素の識別名が付与されているので、変数化部に対応する構文要素となるように、あらかじめ構文解析結果を予測して構文解析を行ない翻訳結果の訳語を得ることができる。

第８図の例では、character が!Nに対応し、最初の!Nの訳は、「キャラクタ」になることがわかる。さらに、ＡＳＣＩＩが２番目の!Nに対応し、

語が「ＡＳＣＩＩ」となる。

（７０３）訳文パターン中の変数化部を（702）のステップで抽出した訳語で置き換える。

第８図では、訳文パターン「この!Nセットは!Nコードと対応が異なる。」に対して、２つの変数化部が存在し、最初の変数化部には「キャラクタ」が、２番目の変数化部には「ＡＳＣＩＩ」がそれぞれ対応して置き換えられる。

第８図の例では、原文パターン中の変数化部の出現順序と訳文パターン中の変数化部の出現順序が同じ順序で対応付けされているが、両者の順序関係が異なる場合には、訳文パターン中の変数化部の記述において、!(m)N と記述することにより、原文パターン中のｍ番目の変数化部!Nに対応することを明記し、変数化部の出現順序を制御することができる。なお、!(m)N の記述において、(m)が省略された場合は、原文パターン中の出現順序と同じ順序関係で対応しているとみなす。さらに、訳文パターン中に、原文パターン中の同一の変数化部に対応する変数化部が複数存在してもよく、その場合には、(m) は省略できないものとする。

（７０４）（７０１）で抽出した変数化部をすべて翻訳処理したかどうかチェックし、未完であれば（７０２）の処理へ移行し、完了であればすべての処理を終了する。

以上の処理結果、翻訳対象テキストに対応する訳文テキストとして、「このキャラクタセットはＡＳＣＩＩコードと対応が異なる。」が最終的に得られることになる。

（２０６）パターン主導型の翻訳処理の結果、訳文の編集が行なわれた場合、（２０９）の処理へ移行する。これは、登録済みの典型対訳パターンが適切でなくて訳文の修正が必要になった場合を考慮して典型対訳パターンを補正するためである。

（２０７）英文テキストを、既に公知となっている翻訳処理装置を用いて翻訳する。翻訳処理装置の実現方式としては、例えば、特開昭58－40684 （自然言語間の自動翻訳方式）に開示された方法を用いることが可能であり説明す省略する。

（２０８）典型パターンの翻訳を行なう。

（２０７）と同様、英文テキストを、既に公知となっている翻訳処理装置を用いて翻訳する。翻訳後、対訳パターンを抽出・登録するために(209)の処理へ移行する。

（２０９）英文テキストとその翻訳結果によって得られた訳文テキストから対訳パターンを抽出し、抽出した対訳パターンを典型対訳ファイル９に格納する。

ここで、対訳事例から対訳パターンを抽出する方式について説明する。対訳パターンの抽出として、新規にパターンを登録する場合と、新規登録後、訳文編集の結果を反映してパターンを補正する場合の２つのケースがある。

以下、第１０図に示す処理フローに従って、それぞれのケースについて説明する。

（１００１）典型対訳ファイル９に、対応する対訳パターンの訳文パターンが登録されているかどうかチェックする。登録されていれば、(1006)の処理へ移行し、対訳パターンを修正する。未登録であれば、（１００２）の処理へ移行し対訳パターンを新規に登録する。

まず、（１００２）～（１００５）の処理ステップにより、新規に対訳パターンを登録する場合について説明する。新規にパターンを登録するのは、典型テキストを翻訳した直後である。テキストの類似度によりクラスタリングされた結果は、類似のテキスト毎にクラスタリングファイル１０にテキスト番号が格納されており、間接的に類似のテキストを参照することができる。各クラスタの典型テキストは、出現順序が最初のテキストであり、第６図（ａ）では、３つのテキストがクラスタリングされており、最初の（Ｅ１－１）が原文典型テキストであることを示す。つまり、（Ｅ１－１）の翻訳処理が終わった段階で、典型対訳パターンを抽出することになる。

（１００２）クラスタリング中の類似文を比較し、類似文の共通部分と、相違部分を識別する。これは、（２０１）で説明したテキスト間の類似度算出方式を利用することによって実現できる。

すなわち、クラスタリング中の任意の２文同士の類似部分の抽出を行ない、さらに同一クラスタリング中の他の文との類似部分の抽出を繰り返していくことにより、クラスタリング中の文の共通部分と相違部分を識別できる。

例えば、第６図（ａ）では、（Ｅ１－１），（Ｅ１－２），（Ｅ１－３）の３つの文について、網かけした部分が相違部分として抽出できる。

（１００３）（１００２）で抽出した相違部分の構文要素を原文典型テキストの翻訳処理での解析結果から決定し、構文要素に対応する構文要素記号を抽出する。第６図（ａ）では、(Ｅ１－１)，（Ｅ１－２），（Ｅ１－３）の３つの文について、破線部で囲んだ構成要素が相違部分である。この部分は、（Ｅ１－１）の翻訳処理結果から名詞句として認識され、構文要素記号!Nを抽出する。

（１００４）原文典型テキスト中に相違部分を該当する構文要素記号で置換し、原文パターンを抽出する。第６図（ａ）の例では、“printer”を名詞句を示す構文要素記号!Nで置き換えることにより、原文パターン（Ｅ１－Ｐ）を抽出する。このようにして、原文パターン中には、変数化部として構文要素記号を持つことになる。

（１００５）原文典型テキスト中の変数化部に対応する訳文典型テキスト中の訳語の部分を、同じ変数化部の要文要素記号で置き換えることにより、対訳パターンの訳文パターンを抽出する。第６図（ａ）の例では、（Ｊ１－１）の「プリンタ」を!Nに置換することにより訳文パターン（Ｊ１－Ｐ）を抽出することができる。

以上のようにして、第６図（ａ）では、対訳パターンとして、（Ｅ１－Ｐ）と（Ｊ１－Ｐ）が抽出できる。

つぎに、（１００６）～（１０１３）の処理ステップにより、登録済みの対訳パターンを補正する場合について説明する。第６図（ｂ）と（ｃ）に示すように、既に対訳パターンが登録されている状況で、パターン主導型の翻訳処理により翻訳した訳文を後編集した場合を具体例として説明する。

（１００６）典型訳文パターンを利用したパターン主導型の翻訳結果の訳文（以下、一次訳と呼ぶ）を、変数化部分に対応した訳出部分と共通部分に対応した訳出部分に区分する。

（１００７）編集した訳文と一次訳を比べて、編集箇所を抽出し、さらに、編集箇所が共通部分と変数化部のいずれに該当するかを区分する。区分する方法としては、（２０１）で説明したテキスト間の類似度算出方式を利用することにより、一次訳と編集した訳文の類似部分と相違部分を識別すればよい。

（１００８）編集箇所が典型対訳パターン中の共通部分に該当するかどうかにより、該当すれば（１００９）へ、該当しなければ（１０１１）へ移行する。該当しない場合、すなわち、編集箇所が典型対訳パターン中の変数化部に該当する場合は、変数化部分に対応する構文要素記号を訳文パターン中に残せばよい。

（１００９）編集箇所に対応する原文の構成要素が原文パターンに包含されているかどうかをチェックし、包含されていれば（１０１２）へ、包含されていなければ（１０１０）へ移行する。このチェックは、編集箇所が典型対訳パターン中の共通部分に該当する場合、編集箇所に対応する原文の構成要素が原文パターンに包含されているかどうかによって２つのケースを考慮しなければならないことに起因する。

例えば、第６図（ｂ）の（Ｊ２－３－１）の編集箇所「文字」に関して、（Ｅ２－３）は（Ｅ２－Ｐ）に包含されるケースであり、編集箇所を訳文パターン中に残すだけでよい。一方、第６図（ｃ）の（Ｊ３－３－１）の編集箇所「日本語」に関して、（Ｅ３－３）は（Ｅ３－Ｐ）に包含されないケースであり、"Japanese"が包含されない構成要素として抽出されるため、訳文パターンと原文パターンの補正が必要になる。

（１０１０）編集箇所に対応する原文中の構成要素を抽出し、さらに原文の解析結果からその構文要素を抽出し該当する構文要素記号を得る。第６図（ｃ）の（Ｅ３－３）は"Japanese"が（Ｅ３－Ｐ）に包含されない構成要素であり、構成要素記号として!Nを抽出する。

（１０１１）編集箇所に対応する構文要素記号を、原文と編集した訳文のそれぞれ対応する文字列の部分と置き換える。

（１０１２）すべての編集箇所について処理を終了したら（１０１３）へ、未終了であれば次の編集箇所の処理に移るべく（１００８）へ移行する。

（１０１３）一次訳と編集訳文を比較し、原文と編集訳文中の変数化部分に該当する箇所で構成要素記号になっていない文字列の部分を対応する構文要素記号に置換する。

第６図（ｂ）では、（Ｊ２－３－１）の「JIS」に対応する部分が変数個部分に該当するため、（Ｊ３－３－１）と（Ｅ３－３）の対応する部分を構文要素記号!Nに置換する。

以上の処理により、登録済みの対訳パターンを補正することができる。

（１０１４）抽出ないし補正した典型対訳パターンを典型対訳パターンファイル９へ格納する。併せて、更新日時をセットする。

以上の処理により典型対訳パターンの抽出・登録および補正を行なうことができる。

（２１０）訳文ファイル６に、訳文テキストを格納する。併せて、更新日時をセットする。

（２１１）全ての英文テキストを翻訳したかどうかチェックし、未完であれば（２０２）の処理へ移行する。完了であれば、すべての処理を終了する。

上記の説明では、本発明による翻訳処理装置の動作について説明した。次に、利用者から見た翻訳処理装置のディスプレイ装置１の表示例について説明する。

利用者は、翻訳対象となる英文テキストを指定した後、ディスプレイ装置１上で必要に応じて翻訳結果を編集することになる。上記の説明では、３つの処理ステップ（２０５）と（２０７）と（２０８）の各翻訳処理の終了後に、利用者の訳文編集の介入を可能とする。以下、本発明による

ディスプレイ装置１上での利用イメージについて第９図を用いて説明する。

第９図（ａ）は、翻訳結果編集モードでのディスプレイ装置１上の表示レイアウトを示す。英文表示エリアと訳文表示エリアを対訳表示するとともに、対訳毎に類似文の有無表示エリアを対訳表示に対応して表示する点に特徴がある。なお、類似文の有無表示エリアは、対訳に対応していれば良く、画面上の左端や右端に表示しても良い。さらに、必要に応じて参照可能なデータの表示エリアとして、典型対訳パターン表示エリアと類似対訳事例表示エリアがある。

類似文の有無表示エリアの表示例を第９図（ｂ）および（ｃ）に示す。原文と訳文の対訳に付随して“Ｐ”がセットされていると、対訳に対応する典型対訳パターン、および類似対訳事例が、第９図（ａ）に示すそれぞれのエリアで参照可能となる。すなわち、利用者の立場からすれば、類似文の有無表示エリアをチェックするだけで参照可能情報の有無を確認できる。

原文と訳文の対訳表示においては、典型対訳パターン中の変数化部分と共通部分に対応する部分を識別表示する。さらに、典型対訳パターンおよび類似対訳事例の表示においても、変数化部分と共通部分に対応する部分を識別表示する。表示の一例を、第９図（ｂ）と第９図（ｃ）に示す。図中、テキストの網かけした部分が変数化部分に該当し、その他の部分が共通部分に該当する。変数化部分は、さらに１対１の対応関係が一目で分かるように色別に表示することも可能である。

また、訳文の表示エリアにおいて、共通部分の表示部分を編集不可能とし変数化部分のみ編集可能とすることにより、訳文の編集操作性を向上することも可能である。これは、パターン主導型の翻訳処理時に、変数化部に対応する部分とそれ以外の部分を識別しておくことにより容易に実現できる。その際、訳文の編集操作時に共通部分に対応する文字列を修正したい場合には、共通部分の編集不可能モードを編集可能モードに変更するための編集変更キーを設定しておき、編集変更キーの押下により編集可能とするようにしてもよい。これにより、共通部分と変数化部分の編集操作を区別することができ、編集箇所が共通部分と変数化部分のいずれに対応するかのチェックが容易に実現できる。また、共通部分の修正のためのキー操作を設けることにより、共通部分の誤修正を回避できるという副次的な効果もある。一方、原文の表示においても同様の区別を行なうことにより、原文が修正された場合、変数化部分に対応する部分の修正のみであればパターン主導型の翻訳処理を適用し、共通部分に対応する部分の修正があれば既存の翻訳処理を適用するといったように、原文の編集対象部分によって、適用する翻訳手段を自動的に切り替えることも容易に実現できる。

次に、類似対訳事例の表示方法について補足説明しておく、類似対訳事例は、クラスタリングファイル１１中の原文番号と訳文番号を参照キーとして原文ファイル５と訳文ファイル６からそれぞれテキストを検索し対訳表示すればよい。その際、複数の類似文の表示の順序の設定方法として、編集対象の原文との類似度の大きい順に並べかえて表示することにより、利用者がより迅速に利用価値の高い類似文を参照できるようにすることが可能となる。また、類似文として抽出されてテキストの中には、お互いに全く一致するテキストの存在も考えられるので、完全一致するテキストは重複表示をしないように事前にチェックした後、お互いに異なる類似文のみ表示するようにすることも可能である。また、編集対象の原文と完全一致する類似文も表示しないように事前にチェックすることが可能である。重複して出現する類似文については、出現頻度を対訳と併せて表示することも効果的である。

さらに、原文ファイル５，訳文ファイル６，クラスタリングファイル１１，典型対訳パターンファイル９中の各データに付随して設定されている更新日時をチェックすることにより、例えば、典型対訳パターンファイル９中のパターンが補正された場合、補正された更新日時以前に訳文ファイル６中に格納された類似文の訳文テキストのみ翻

訳編集画面上に対訳表示することも可能である。

以上の実現方法および表示方法は、上記の実施例の説明から容易に類推できるものであり、本発明の要旨を逸脱しない範囲で種々変更して実施することが可能である。

以上、本発明による一実施例を説明した。

本発明の適用効果として、上記手段により本発明を適用した翻訳処理装置では、既に翻訳した文書の対訳を利用することにより、類似文を翻訳する場合、類似典型パターンと異なる要素の部分のみ翻訳すればよく、翻訳処理を軽減することができる。さらに、入力文書に対しても予め類似文毎にクラスタリングしておくことにより入力文全てを機械翻訳処理の対象とする必要がなくなり、翻訳処理効率が向上する。また、従来の翻訳処理装置では、類似文でも構成要素が微妙に異なっただけで入力文の構文解析結果が微妙に異なったり、解析処理を失敗したりといった問題が発生していたが、本発明により、類似の文の翻訳精度は均一に保てるというメリットも生じる。

実施例2

別の実施例として、英語から日本語への翻訳を行なう翻訳処理装置について述べる。装置の構成は実施例1と同様とする。

以下、本発明による翻訳処理の流れを第11図に示すフローに従って具体的に説明する。

（1101）英文ファイル5に格納されている翻訳対象英文について、全ての英文テキスト同士の類似度を評価したのち、英文テキストのクラスタリングを行なう。これは、実施例1のステップ（201）で説明したテキスト間の類似度を利用した英文テキストのクラスタリング処理方式を利用すればよく、クラスタリングファイル10にクラスタ毎に類似の英文テキストの文番号を蓄積する。これらの類似の英文中、出現順序が最初の英文がクラスタの典型テキストとなっている。

（1102）原文ファイル5から翻訳対象の英文テキストを出現順に抽出する。

（1103）英文テキストが典型テキストかどうかチェックし、典型テキストであれば（1104）、典型テキストでなければ（1109）の処理へ移行する。典型テキストのチェックは、原文ファイル5中のクラスタリング番号によりアクセスしたクラスタリングファイル11中の原文番号の登録順序が1番目かどうかをチェックすればよい。

（1104）典型テキストの翻訳を行なう。さらに、翻訳結果の編集を行なう。ここでの、翻訳・編集処理は、公知となっている機械翻訳処理装置を使用すればよい。翻訳処理装置の実現方式としては、例えば、特開昭58－40684（自然言語間の自動翻訳方式）に開示された方法を用いることが可能であり説明は省略する。

（1105）典型テキストの翻訳結果と編集結果、さらに典型テキストの類似テキストを利用して、典型対訳パターンを抽出し、典型対訳ファイル9に登録する。ここで、典型対訳パターンの抽出方法については、実施例1のステップ（209）と同様の方式により実現できるので説明は省略する。

（1106）典型テキストの類似テキストをクラスタリングファイル11の原文番号を参照して原文ファイル5より順次抽出する。

（1107）抽出した類似テキストを、典型対訳ファイル9に登録してある典型原文パターンと典型訳文パターンの対応関係を利用して、パターン主導型の翻訳処理を行なう。

（1108）典型テキストの類似テキストすべてについて抽出および翻訳を終了したかどうかチェックし、未終了であれば（1106）、終了であれば（1111）へ移行する。

（1109）抽出した英文テキストが類似文を持つかどうかをチェックする。チェック方法は、原文ファイル5中の英文テキストに対応してクラスタリング番号が存在するかどうかをチェックすればよい。チェックの結果、類似文を持てば、既に（1106）と（1107）で翻訳済みであるので（1102）へ移行し次の英文テキストの処理に移る。類似文を持たなければ（1110）へ移行する。

（1110）英文テキストを翻訳する。翻訳処

理は、公知となっている機械翻訳処理装置を使用することが可能であり説明は省略する。

（１１１１）翻訳結果の訳文を編集する。ここで、訳文の編集対象として、典型テキスト翻訳時の類似文の翻訳結果も併せて編集可能とする。この点について、第１２図を用いて表示イメージを含めて説明する。

第１２図（ａ）は、原文ファイル５中の英文テキストを翻訳編集する場合の画面例である。図中、左半分が原文表示エリア、右半分が訳文表示エリア、更に中央に類似文の有無表示エリアがありエリア中の“Ｐ”が類似文の存在を示している。今、２番目の文で“Japanese set has a different mapping from the JIS code.”が典型テキストとなる文を翻訳編集した直後の状況を考える。典型テキストには類似テキストが存在し、ステップ（１１０６）から（１１０８）により類似テキストの翻訳が行なわれるので、典型テキストの翻訳編集結果と類似文の翻訳結果を第１２図（ｂ）に示すような翻訳編集画面の型式で表示する。すなわち、翻訳編集画面の上部に典型テキストの対訳を表示し、その下の編集エリアに類似文の翻訳結果をまとめて表示する。その際、典型対訳パターン中の変数化部に対応する部分を識別表示しておく。翻訳装置に利用者は、典型テキストの翻訳結果を参考にしながら、類似テキストの訳文を編集することが出来る。類似文の対訳表示エリアは、テキストが多く一画面に収まらないときには適宜画面をスクロールする。類似文の編集が終了した段階で、翻訳編集画面は、第１２図（ｃ）に示すように典型テキストの類似文に関してのみ翻訳編集を完了したことを反映して、対訳表示中類似テキストのみ部分的に訳文が表示されることになる。

（１１１２）原文ファイル５中の翻訳対象テキストすべてについて処理が終了すれば翻訳処理を完了し、未完であれば（１１０２）へ移行し処理を続行する。

以上、本発明による別の一の実施例を説明した。

上記手段により、本発明の適用した翻訳処理装置では、翻訳対象となる入力文書に対して予め類似文毎にクラスタリングしておき、クラスタリング中の最初のテキストの翻訳編集と同期して、類似のテキストについても翻訳編集を行なうことにより、翻訳校正作業が効率良くできると共に、類似の文の翻訳精度を均一に保てるという効果が生じる。

また、上記の実施例では、翻訳処理装置を例にとって説明したが、ワードプロセッサのような文書処理装置に対しても本発明の適用は可能である。すなわち、まとまりのある文書を作成・編集する場合に、文書中に出現する類似の文を抽出・管理し、類似文の表現を統一するための手段として使用したり、逆に類似の表現の多用をチェックする手段として使用する等の用途が考えられる。このような手段は本発明の要旨を逸脱しない範囲で種種変形して実施することが可能である。

実施例３

別の実施例として、英語から日本語への翻語を行なう翻訳処理装置について述べる。装置の構成は実施例１と同様とする。

本実施例では、第１図の典型対訳パターンファイル９中に既にパターンが登録してあることを前提にする。これは、同じ種類の文書、例えば、コンピュータマニュアルの世界では、ＰＬ／ＩやＣＯＢＯＬ等の言語プロセッサの使用マニュアルが多々存在するが、ＰＬ／Ｉの使用マニュアルを翻訳するときにＣＯＢＯＬの使用マニュアル中の表現と同一乃至類似の表現が使用されることが多い。そこで、翻訳対象の文書と同類の文書で既に対訳が存在していれば、その対訳から抽出される典型対訳パターンを利用して翻訳作業効率を向上しようという考え方に基づくものである。

以下、本発明による翻訳処理の流れを第１６図に示すフローに従って具体的に説明する。

（１６０１）原文ファイル５中の翻訳処理英文テキストを出現順に順次取り出す。

（１６０２）抽出した英文テキストについて、典型対訳パターンファイル９中の原文パターンと類似な関係になる対訳パターンを抽出する。ここで、類似の対訳パターンの抽出方法を第１３図に

示すフローに従って説明する。

（１３０１）メモリ４中に設定した類似文保持テーブルをクリアする。類似文保持テーブルには、典型対訳パターンファイル９から抽出した類似な対訳パターンに対応するパターン番号と類似度をペアにして複数個格納できるようにしたものであり、テーブルクリアではこれらのテーブルの値をゼロにする。類似文保持テーブルには、以下のステップにより、類似度の大きい順にパターン番号と類似度を類似度の降順に限定個数（例えば、５個）登録していく。

（１３０２）典型対訳パターンファイル９から原文パターンを順次読み出す。

（１３０３）翻訳対象の英文テキストと原文パターンの類似度を算出する。２つのテキスト間の類似度の算出方法は、実施例１のステップ（201）で説明した方法を用いればよい。

（１３０４）抽出した類似度と類似文保持テーブルに既に登録済みの類似な対訳パターンの候補の類似度とを比較し、抽出して類似度の方が大きければ、類似度保持テーブルにパターン番号と類似度を降順になるように追加し、最小の類似度をもつパターン番号を削除する。抽出した類似度の方が小さければ、何もしない。

（１３０５）典型対訳パターンファイル９中のすべての原文パターンを読み出したかどうかチェックし、未完であれば（１３０２）へ、完了であれば（１３０６）へ移行する。

（１３０６）類似文保持テーブルには、類似度の大きい順に典型対訳ファイル９中の対訳パターンへのアクセスのためのパターン番号が格納されている。このパターン番号により対応する原文パターン，訳文パターン，クラスタリング番号を抽出する。さらに、クラスタリング番号からクラスタリングファイル１１中の類似のテキストの原文番号と訳文番号を抽出し、類似の対訳データを抽出する。

以上の処理により、翻訳対象の英文テキストに類似な対訳パターンと対訳文データが抽出できる。

（１６０３）類似のパターンがあるかどうかを類似文保持テーブルによりチェックする。あれば（１６０４）、なければ（１６０６）へそれぞれ移行する。

（１６０４）類似の対訳パターンの中で類似度が最大のものを利用してパターン主導型の翻訳処理を行なう。これは、実施例１で説明した方式と同じであり、説明は省略する。

（１６０５）パターン主導型の翻訳処理結果の訳文を校正する。その際、（１６０２）のステップで抽出した複数の類似な対訳パターンさらに類似な対訳データを対訳表示編集画面上で表示し利用することにより、訳文の校正が出来るようにする。すなわち、対訳表示編集画面の一例として、第９図に示すよう、翻訳対象文の翻訳結果表示エリア，類似な典型対訳パターンの表示エリア，類似な対訳データの表示エリアを識別して表示することが出来る。これらの表示エリアは、必要に応じて表示参照が出来るようにしてもよい。

また、類似な典型対訳パターンおよび対訳データの表示の際に、翻訳対象テキストとの類似度を類値で示したり色別表示するなど明示的に表示してもよい。

（１６０６）既に公知となっている翻訳処理装置を用いて翻訳対象の英文テキストを翻訳する。

（１６０７）（１６０６）での翻訳結果の訳文を校正する。

（１６０８）校正済みの訳文を訳文ファイル６へ格納する。

（１６０９）原文ファイル５中の翻訳対象のすべての英文テキストを翻訳処理したかどうかチェックし、未終了であれば（１６０１）へ移行し、終了であればすべての処理を完了する。

以上、本発明による別の一実施例を説明した。

本発明の適用効果として、類似の文書の典型対訳パターンファイルを予め複数個用意しておくことにより、新たな文書の翻訳の際に、最も類似と思われる文書の典型対訳パターンファイルを選択利用して効率良く翻訳および訳文の校正作業を行なうことができる。

新たな文書を翻訳する際に、複数の典型対訳パ

ターンファイルの中からどのファイルを利用すればよいかという問題については、予め翻訳対象テキストのクラスタリング結果の原文の典型パターンと各典型対訳パターンファイル中の原文パターンとの類似度のチェックを行ない、類似文が最も多く含まれるファイルを利用する方式も可能である。また、翻訳時に利用する典型対訳パターンファイルを1つに限定する必要はなく、複数のファイルを利用するようにしても良い。これらは、本発明の要旨を逸脱しない範囲で種々変形して実施できるものである。

実施例4

別の実施例として、実施例2で説明した英語から日本語への翻訳を行なう翻訳処理装置における典型対訳パターンファイル9の作成方法について述べる。装置の構成は実施例2と同様とする。

実施例2では、典型対訳パターンファイル9内のパターンは翻訳処理を行なう前に作成済みであることを前提に説明したが、本発明の考え方を適用することにより典型対訳パターンを効率的に作成・保守することが可能である。以下、典型対訳パターンの保守方法について説明する。

本実施例での典型対訳パターンの保守方法は、既に登録済みの典型対訳パターンに対して、パターン間の類似性を利用してパターンを再クラスタリングすることにより、パターン数の軽減、さらには複数の典型対訳パターンファイル中のパターンデータを統合することを目指すものである。以下、第14図のフローに従い、典型対訳パターンのマージ方法について説明する。

（1401）典型対訳パターンファイル9中の任意の2つの原文パターンについて類似度を算出する。2つのパターン間の類似度の算出方法は、実施例1のステップ（201）で説明した2つのテキスト間の類似度算出方法と同様の方法を用いることが可能である。

（1402）算出した類似度が予め設定した閾値（例えば、0.8）より大きいかどうかチェックし、大きければ（1403）、小さければ(1404)へ移行する。

（1403）2つのパターンの類似度が大きい、すなわち類似性があるので、2つのパターンをマージする。ここで、2つのパターンのマージ方法については、実施例1のステップ（1002）～（1005）で説明した複数のテキストからのパターン抽出方法を適用すればよい。一例として、第15図に示す2つの典型対訳パターン（E1－P）（J1－P）と（E2－P）（J2－P）をマージする場合、（E1－P）と（E2－P）の比較によりアンダラインを引いた!Nの部分のみ異なり残りの部分は共通となる。従って、パターンのマージ結果は（E1－P－1）（J1－P－1）に示すように網掛けした〈!N〉が任意の構文要素として省略可能なパターンとなる。ここで、〈と〉で囲んだ構文要素記号は省略可能なことを示す。

典型対訳パターンファイル9中には、マージ結果の新たな対訳パターンを登録し、マージの対象となった2つの対訳パターンは削除する。

（1403）典型対訳パターンファイル9中のすべての原文パターンの組合せについて類似度を算出したかどうかをチェックし、未終了であれば（1401）へ移行し、終了であればすべての処理を完了する。

処理完了後、典型対訳パターンファイル9中には、処理前のパターンの数よりも少ないか等しい数のパターンが登録されることになる。

上記の説明では、典型対訳パターンファイルを1つに限定した場合について説明したが、複数の典型対訳パターンファイルから1つの典型対訳パターンファイルにパターンをマージする場合も同様の処理方法が適用できる。

以上、本発明による別の一実施例を説明した。

本発明の適用効果として、典型対訳パターンを自動的に整理・統合することが可能となる。すなわち、翻訳対象の入力原文の増加に伴って典型対訳パターンも一般的に増えていくと考えられる。そのような場合の典型対訳パターンファイルの保守管理方法として、典型対訳パターン中の類似のパターンを統合することにより、典型対訳パターンファイルを保守することが可能となる。さらに、

典型対訳パターンファイル中の登録上の制限として類似文の数がある閾値以上のパターンに限定するといったような制約を設けてもよい。これらは、本発明の要旨を逸脱しない範囲で種々変形して実施できるものである。

実施例5

別の実施例として、実施例2で説明した英語から日本語への翻訳を行なう翻訳処理装置における典型対訳パターンファイル9の作成方法について述べる。装置の構成は実施例2と同様とする。

実施例1では、典型対訳パターンファイル9内のパターンを、翻訳処理実行時に翻訳構成結果を利用して作成する方法について説明した。また、実施例4では、既存の典型対訳パターンファイル9中のパターン情報をマージという手段により整理・統合する方法について説明した。本実施例での基本的な考え方は、既に存在する対訳事例を利用することにより、対訳事例から原文と訳文のそれぞれについて類似度によるクラスタリング処理を行ない、クランスリング結果を利用して典型対訳パターンを新規に作成する方法である。

以下、典型対訳パターンの作成方法について第17図のフローに従って説明する。

（1701）対訳事例ファイル11に格納されている原文テキストと訳文テキストについて、それぞれ類似度によるクラスタリングを行なう。類似度によるクラスタリング方法は、実施例1のステップ（201）で説明した方法を利用すればよい。

対訳事象ファイル11には、原文ファイル5と訳文ファイル6のテキストに該当するデータがペアに格納されているものとする。対訳事例ファイル11の代わりに、原文ファイル5と対応する翻訳校正済みの訳文ファイル6のデータを利用してもよい。

原文テキストと訳文テキストのそれぞれについてクラスタリング処理を行なうことにより、第18図に示すように原文のクラスタと訳文のクラスタさらに各クラスタ中のテキストの対応関係が抽出できる。

（1702）抽出した原文と訳文の対応関係にあるクラスタについて、それぞれのクラスタ中のテキストの数が等しく、さらにテキストの対応関係がとれているかどうかチェックする。テキストの数が等しく対応関係がとれていれば(1705)へ、それ以外は（1704）へ移行する。

ここで、原文と訳文のクラスタの関係は第18図に示すようになる。図中、原文がクラスタAは（＊1，＊2，＊3，＊4）の各テキストがクラスタリングされているが、対応する訳のクラスタA2では（＊1，＊21，＊4）の各テキストがクラスタリングされており、対応関係がとれないため（1704）の処理へ移行することになる。一方、原文のクラスタBと訳文のクラスタBではお互いに（＊7，＊10，＊26）の各テキストがクラスタリングされており、対応関係がとれているため（1703）の処理へ移行することになる。

（1703）対応関係のとれた原文と訳文のクラスタから対訳パターンを抽出し、典型対訳ファイル9に登録する。併せて、原文ファイルに原文テキストを、訳文ファイルに訳文テキストを格納し、さらにクラスタリングファイル11に対応付けのためのデータを格納しておく。ここで、原文と対訳のクラスタからの対訳パターンの抽出方法は、実施例1のステップ（1002）～(1005)で説明した方法を用いることが可能である。また、ファイル間の関係についても実施例1での扱いと全く同様に管理すればよい。

（1704）すべての原文と訳文のクラスタについて処理を終了したかどうかをチェックし、未終了であれば次のクラスタを抽出して(1702)へ移行する。処理終了であればすべての処理を完了する。

以上、本発明による別の一実施例を説明した。

本発明の適用効果として、原文と対応する訳文を対応事例として再利用することにより、典型対訳パターンを自動的に抽出することが可能となる。すなわち、翻訳結果の対訳事例を新たな文書の翻訳の際に積極的に利用することが可能となる。対

訳データを翻訳処理結果の一過性のデータで閉じさせるのではなく、翻訳処理のためのパターンデータとして有効なデータベースとして活用させることが可能となる。

〔発明の効果〕

以上説明したごとく本発明を適用した翻訳装置では、既に翻訳した文書の対訳を利用することにより、類似文を翻訳する場合、類似典型パターンと異なる要素の部分のみ翻訳すればよく、翻訳処理を軽減することができる。さらに、入力文書に対しても予め類似文毎にクラスタリングしておくことにより入力文全てを機械翻訳処理の対象とする必要がなくなり、翻訳処理効率が向上する。また、従来方式では、類似文でも構成要素が微妙に異なっただけで入力文の構文解析結果が微妙に異なったり、解析処理に失敗したりといった問題が発生していたが、本発明により、類似の文の翻訳精度は均一に保てるというメリットも生じる。

さらに、翻訳対象となる入力文書に対して予め類似文毎にクラスタリングしておき、クラスタリング中の最初のテキストの翻訳編集と同期して、類似のテキストについても翻訳編集を行なうことにより、翻訳校正作業が効率良くできると共に、類似の文の翻訳精度を均一に保てるという効果が生じる。

また、類似の文書の典型対訳パターンファイルを予め複数個用意しておくことにより、新たな文書の翻訳の際に、最も類似と思われる文書の典型対訳パターンファイルを選択利用して効率良く翻訳および訳文の校正作業を行なうことができる。

本発明の他の適用効果として、原文と対応する訳文を対訳事例として再利用することにより、典型対訳パターンを自動的に抽出することが可能となる。すなわち、翻訳結果の対訳事象を新たな文書の翻訳の際に積極的に利用することが可能となる。対訳データを翻訳処理結果の一過性のデータで閉じさせるのではなく、翻訳処理のためのパターンデータとして有効なデータベースとして活用させることが可能となる。その結果、翻訳結果を再利用することにより翻訳処理のためのパターンデータを逐次自動的に増補していくことが容易に実現できるため、翻訳装置の利用者は翻訳規則の調整を行わずとも翻訳精度を向上させることが可能となる。

併せて、典型対訳パターンファイルの保守管理方法として、典型対訳パターン中の類似のパターンを統合することにより、典型対訳パターンファイルを保守することが可能となる。その結果、翻訳対象の入力原文の増加に伴って典型対訳パターンが増えても、有効な典型対訳パターンを自動的に調整・統合することが実現できる。

実施例の説明では、英語から日本語への2か国語間の翻訳処理装置を例にとって説明したが、典型対訳パターンを2か国語間のパターンからNか国語（N＞2）のパターンに拡張することにより3か国語以上の翻訳装置へ適用することは、本発明の要旨を逸脱しない範囲で実現できるものである。

また、翻訳装置以外にも本発明は適用可能であり、本発明を適用した文書処理装置においては、文書の編集操作時に、文書中の類似文を確認することが容易になるため、マニュアルなどの編集操作において、表現のバラツキを防止し、表現上統一された文書の校正処理効率が向上する。すなわち、まとまりのある文書を作成・編集する場合に、文書中に出現する類似の文を抽出・管理し、類似文の表現を統一するための手段として使用したり、逆に類似の表現の多用をチェックする手段として使用する等の用途が考えられる。

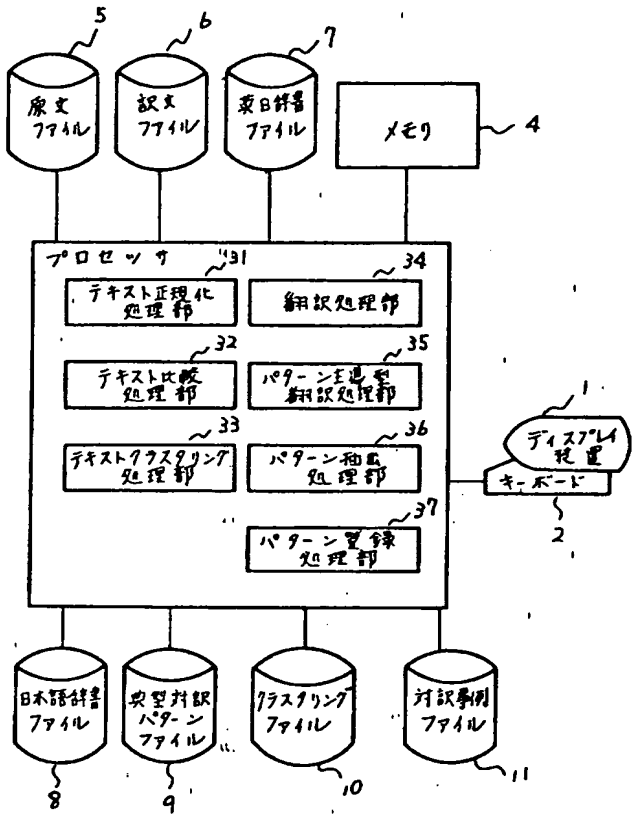
4．図面の簡単な説明

第1図は、本発明の一実施例の構成を示すブロック図、第2図は、本発明の一実施例の動作を説明するための処理フロー図、第3図は、第1図中のファイル間の関係を説明する説明図、第4図は、本発明の一実施例の動作を説明するための処理フロー図、第5図，第6図は、本発明の一実施例の動作を説明するための説明図、第7図は、本発明の一実施例の動作を説明するための処理フロー図、第8図，第9図は、本発明の一実施例の動作を説明するための説明図、第10図，第11図は、本

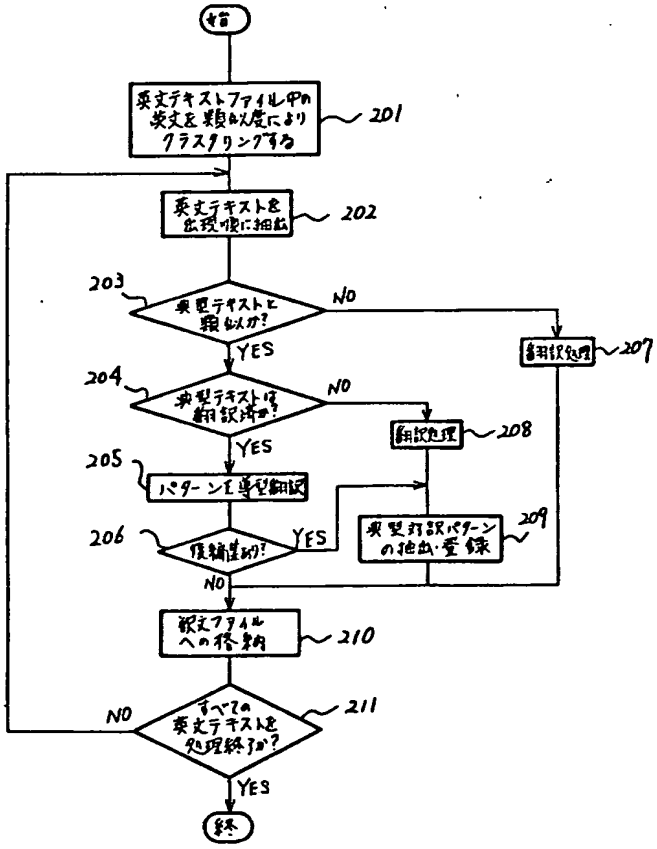
発明の一実施例の動作を説明するための処理フロー図、第１２図は、本発明の一実施例の動作を説明するための説明図、第１３図，第１４図は、本発明の一実施例の動作を説明するための処理フロー図、第１５図は、本発明の一実施例の動作を説明するための説明図、第１６図，第１７図は、本発明の一実施例の動作を説明するための処理フロー図、第１８図は、本発明の一実施例の動作を説明するための説明図をそれぞれ示す。

１…ディスプレイ等の出力装置、２…キーボード等の入力装置、３…プロセッサ、４…メモリ、５…原文ファイル、６…訳文ファイル、７…英日辞書ファイル、８…日本語辞書ファイル、９…典型対訳パターンファイル、１０…クラスタリングファイル、１１…対訳事例ファイル、３１…テキスト正規化処理部、３２…テキスト比較処理部、３３…テキストクラスタリング処理部、３４…翻訳処理部、３５…パターン主導型翻訳処理部、３６…はパターン抽出処理部、３７…パターン登録処理部。

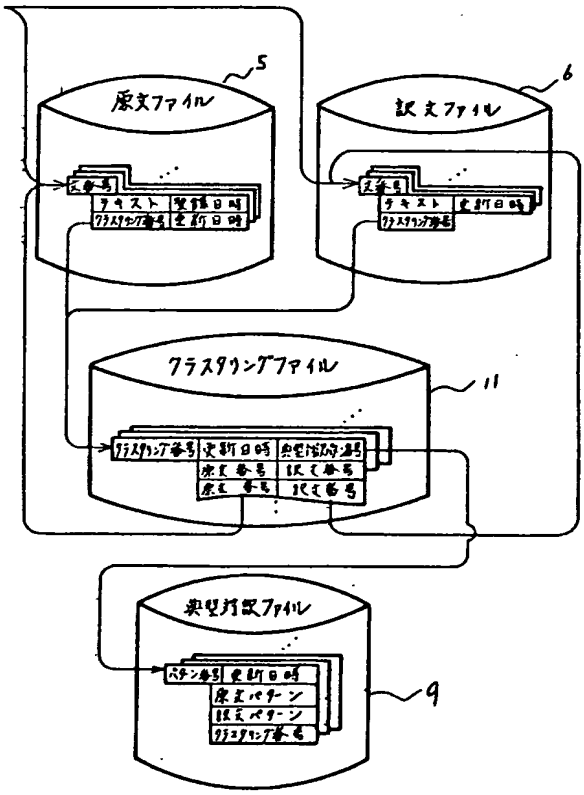
第１図



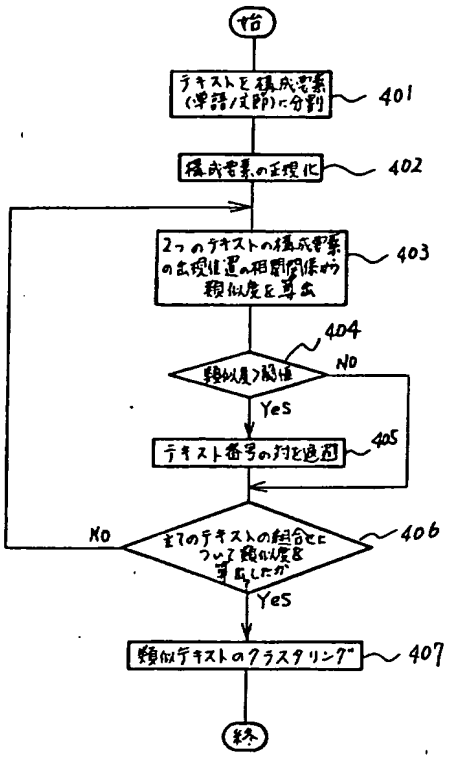
第２図



第３図

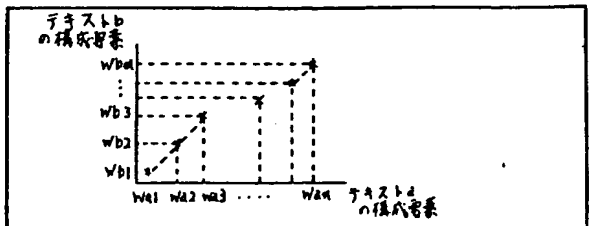


第 4 図



第 5 図

(a)



類似度 = (2×単語の連続・設定個数 − 誤単語数 − 不一致単語数) / 総単語数

(注:総単語数は文末の句読点を除く)

(b)

(I-1) Followings are the explanation of printer configuration
「次に、プリンタの構成について説明する」

(I-2) Followings are the explanation of interface configuration
「次に、インタフェースの構成について説明する」

(I-3) Followings are the explanation of system configuration
「次に、システムの構成について説明する」

(I-1)と(I-2)の類似度 = $\frac{2\times4+6-2}{7+7}$ = 0.86

(I-1)と(I-3), (I-2)と(I-3)の類似度も同様

(c)

(J-1) This set has a different mapping from the KEIS code
「この__セットはKEISコードと対応が異なる」

(J-2) This Japanese set has a different mapping from the JIS code.
「この日本語セットはJISコードと対応が異なる」

(J-1)と(J-2)の類似度 = $\frac{2\times5+8-3}{10+11}$ = 0.91

第 6 図

(a)

(E1-1) Followings are the explanation of printer configuration.
(E1-2) Followings are the explanation of interface configuration.
(E1-3) Followings are the explanation of system configuration.
パターン化
(E1-P) Followings are the explanation of ~~printer~~ configuration.

(J1-1) 次に、プリンタの構成について説明する.
(J1-2) 次に、インタフェースの構成について説明する.
(J1-3) 次に、システム__の構成について説明する.
パターン化
(J1-P) 次に、~~printer~~ の構成について説明する

(b)

(E2-1) This Set has a different mapping from the KEIS code.
(E2-2) This Set has a different mapping from the ASCII code.
パターン化
(E2-P) This Set has a different mapping from the ~~printer~~ code.
(E2-3) This Set has a different mapping from the JIS code

(J2-1) このセットはKEISコードと対応が異なる.
(J2-2) このセットはASCIIコードと対応が異なる
パターン化
(J2-P) このセットは ~~printer~~ コードと対応が異なる
パターン翻訳
(J2-3) このセットは JIS コードと対応が異なる
訳文後編集
(J2-3-1) この文字セットは ~~printer~~ コードと異なる対応をしている
パターン補正
(J2-P-1) この文字セットは ~~printer~~ コードと異なる対応をしている.

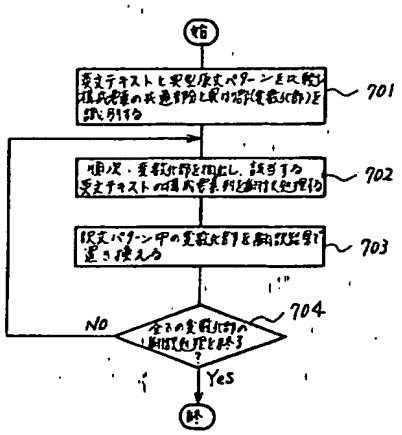
第 6 図

(c)

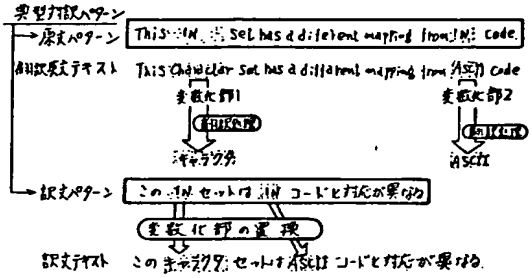
(E3-1) This Set has a different mapping from the KEIS code.
(E3-2) This Set has a different mapping from the ASCII code.
パターン化
(E3-P) This Set has a different mapping from the ~~printer~~ code.
(E3-3) This ~~Japanese~~ Set has a different mapping from the JIS code.
パターン補正
(E3-P-1) This ~~printer~~ Set has a different mapping from the ~~printer~~ code.

(J3-1) このセットはKEISコードと対応が異なる.
(J3-2) このセットはASCIIコードと対応が異なる.
パターン化
(J3-P) このセットは ~~printer~~ コードと対応が異なる.
パターン翻訳
(J3-3) このセットは JIS コードと対応が異なる
訳文後編集
(J3-3-1) この日本語セットは ~~printer~~ コードと対応が異なる.
パターン補正
(J3-P-1) この ~~printer~~ セットは ~~printer~~ コードと対応が異なる

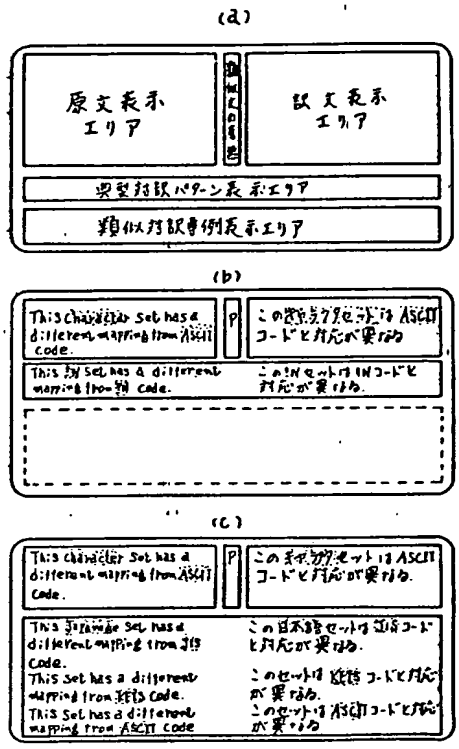
第7図



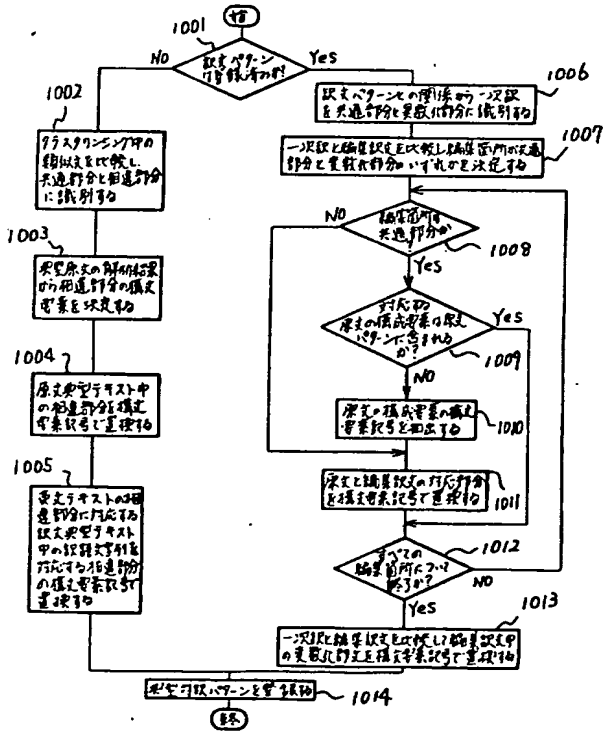
第8図



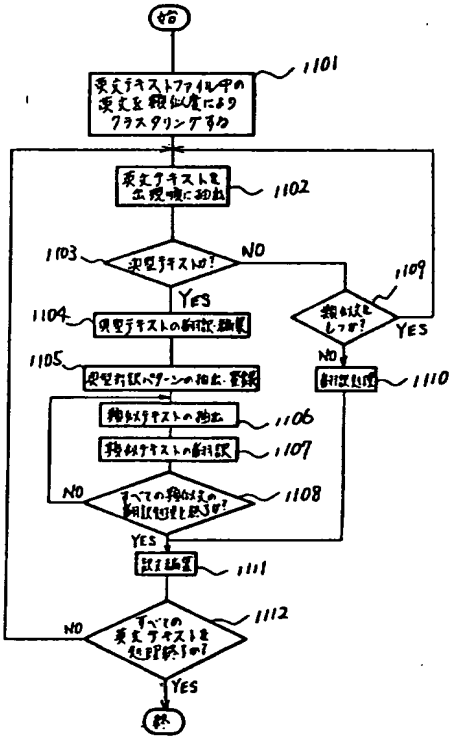
第9図



第10図

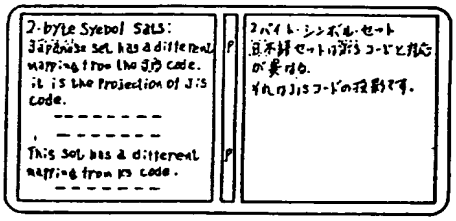


第11図

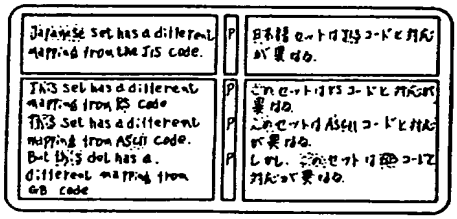


第 12 図

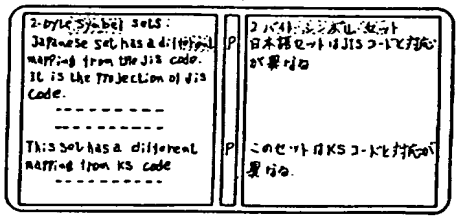
(a)



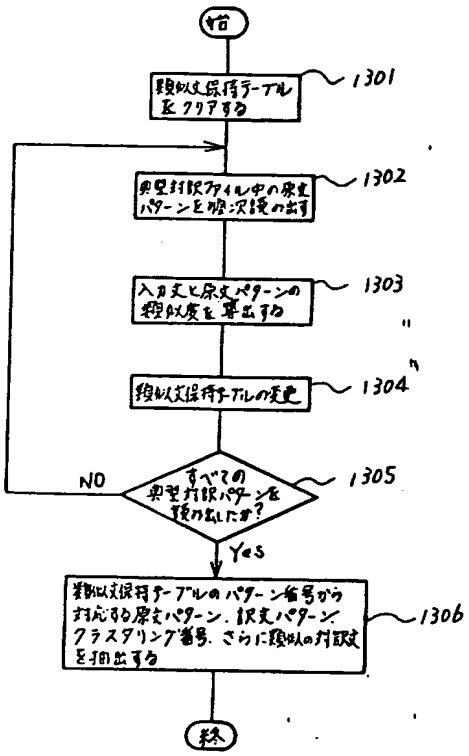
(b)



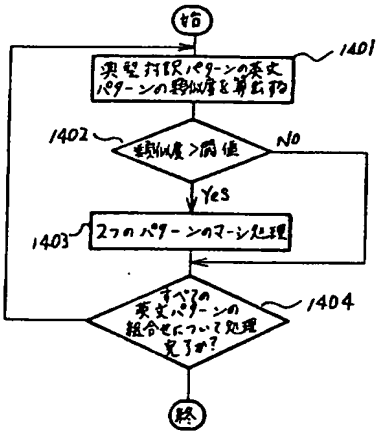
(c)



第 13 図



第 14 図



第 15 図

(E1-P) This Set has a different mapping from the !N Code.

(J1-P) このセットは !N コードと対応が異なる。

(E2-P) This !N Set has a different mapping from the !N Code.

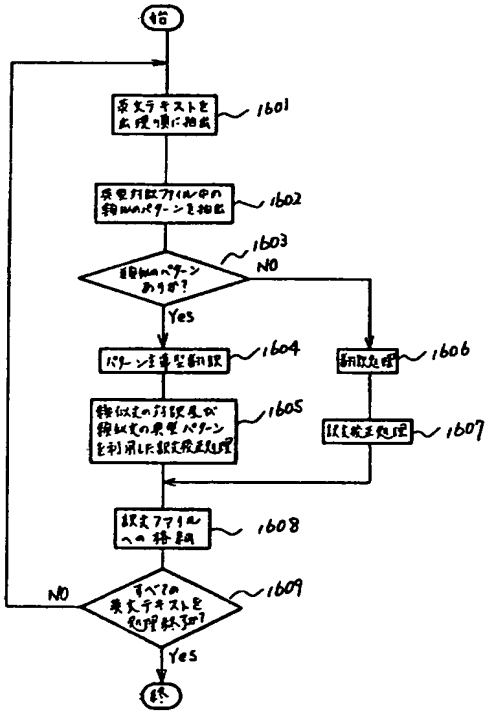
(J2-P) この !N セットは !N コードと対応が異なる

2つのパターンのマージ

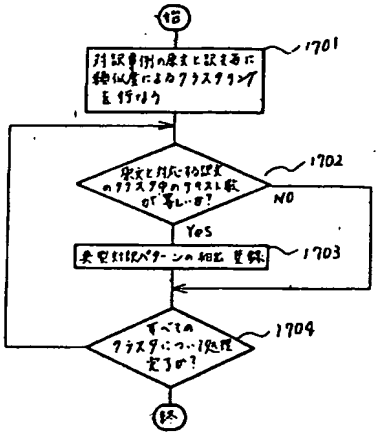
(E1-P1) This [illegible] Set has a different mapping from the !N code.

(E1-P1) この [illegible] セットは !N コードと対応が異なる。

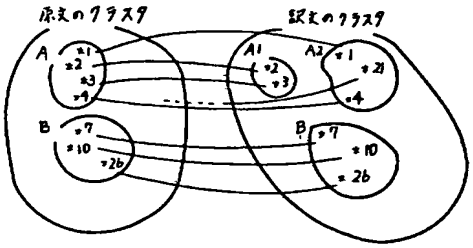
第 16 図



第17図


開始
対訳事例の原文と訳文毎に類似度によるクラスタリングを行なう
1701
原文と対応する原文のクラスタのクラスタ数が等しいか?
1702
NO
Yes
未定対応パターンの組込み登録
1703
すべてのクラスタについて処理完了か?
1704
終了

第18図


原文のクラスタ
訳文のクラスタ
A
B
A1
A2
B

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☑ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☑ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.